CS 134: Networks — Problem set 4

Prof. Yaron Singer — Due: **Wednesday 3/1/2017**

1. Centrality (Jackson 2.2.4). The *degree centrality* of a node is its degree divided by $n-1$ (where n is the number of nodes in the network). For example, the degree centrality of node d in network A (see appendix) is $\frac{3}{5}$.

a. Compute the degree centrality of each of the nodes in network B.

Closeness Centrality. The most basic closeness centrality measure is the inverse of the average distance between i and all other nodes j in the network[1]. To illustrate, let us compute the closeness centrality of node b in the following network

$$(\{a, b, c\}, \{ab, ac\}).$$

- The distance between node b and node a is 1.
- The distance between node b and node c is 2.

The average of these distances is $\frac{3}{2}$, so the closeness centrality of node b in this network is $\frac{2}{3}$.

b. Compute the closeness centrality of node d in network

$$(\{a, b, c, d\}, \{ab, ac, ad, bc, cd\}).$$

2. Betweenness Centrality. Let $P_i(k, j)$ denote the number of geodesics (or shortest paths) between nodes k and j that node i lies on, and let $P(k, j)$ be the total number of geodesics between nodes k and j. We can estimate how important node i is in terms of connecting nodes k and j by looking at the fraction $\frac{P_i(k,j)}{P(k,j)}$ of shortest paths between nodes k and j that involve node i. Averaging this ratio across all pairs of nodes (aside from node i) gives us the *betweeness centrality* of node i. To illustrate, let us compute the betweenness centrality of node c in the following "circle network"

$$(\{a, b, c, d\}, \{ab, bc, cd, da\}).$$

- None of the shortest paths between node a and node b involve node c, so $\frac{P_c(a,b)}{P(a,b)} = 0$.
- None of the shortest paths between node a and node d involve node c, so $\frac{P_c(a,d)}{P(a,d)} = 0$.
- One of the two shortest paths between nodes b and node d involve node c, so $\frac{P_c(b,d)}{P(b,d)} = \frac{1}{2}$.

Averaging out these three ratios, we find that the betweenness centrality of node c in this network is $\frac{1}{6}$.

Compute the betweenness centrality of each of the nodes in the network $(\{a, b, c, d, e, f\}, \{ab, bc, cd, ce, df, ef\})$.

[1] There are various conventions for handling networks that are not connected, as well as other possible measures of distance, which lead to a family of closeness measures, but for now we won't get into this.

3. Programming: Verifying the Friendship Paradox. Using the eight sample directed social networks below, courtesy of SNAP (https://snap.stanford.edu/data), for each social network, compute the value of

$$\frac{\frac{1}{|G|}\sum_{n\in G}(deg(n))}{\frac{1}{|G|}\sum_{n\in G}(\frac{1}{|N(n)|}\sum_{m\in N(n)}deg(m))}$$

with the average being taken across all nodes n in the network G, and where $N(n)$ is the neighborhood of node n.

(That is, calculate the value of (average node degree/average average degree of node's neighbors) in the graph.)

If the networks are too large, consider a reasonable (what constitutes "reasonable" is up to your judgment, as long as you provide a textual defense in your submission) subsample of the nodes.

Keep in mind that the graph is directed, only consider outgoing edges when computing degree, and only consider neighbors to whom node n has an outgoing edge to that neighbor. If a node has out-degree of 0, then assume the average degree of its neighbors is 0.

Compute the value above for each of the following social networks:

- https://snap.stanford.edu/data/com-Orkut.html
- https://snap.stanford.edu/data/com-LiveJournal.html
- https://snap.stanford.edu/data/soc-Slashdot0811.html
- https://snap.stanford.edu/data/com-DBLP.html
- https://snap.stanford.edu/data/email-Enron.html
- https://snap.stanford.edu/data/com-Youtube.html
- https://snap.stanford.edu/data/soc-Epinions1.html
- https://snap.stanford.edu/data/wiki-Talk.html

You should use the first dataset in the 'Files' table on each page. Note that some of these graphs are directed, and some are undirected, as indicated by 'ungraph', which you should account for when building the graphs.

As a **bonus question (just for fun, no extra credit)**, for each network, plot the average of $\frac{deg(n)}{\frac{1}{|N(n)|}\sum_{m\in N(n)}deg(m)}$ against sample size (number of nodes n in the subsample). That is, for samples of varying sizes, compute this average for each sample on the y axis, and plot the sample size on the x axis.

You should also turn in to Canvas a script named '*Lastname_Firstname_ps4.py*', with your name substituted in place of '*Lastname_Firstname*', in addition to the rest of your problem set, and the appropriate extension for the language you're using. If you are using python, see this Piazza link (post 269) for expectations on how to submit your code (specifically: we'll no longer accept python notebooks in submission, though they're fine to use while solving the problem).

Figure 1: Network A

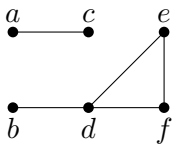

Figure 2: Network B